Overview and Investigation of Algorithms for the Information Bottleneck Method

Shayan Hassanpour, Dirk Wübben, Armin Dekorsy
Department of Communications Engineering
University of Bremen, 28359 Bremen, Germany
Email: {hassanpour, wuebben, dekorsy}@ant.uni-bremen.de

***Abstract*—Lossy data compression has been studied under the celebrated Rate-Distortion theory which provides the compression rate in order to quantize a signal without exceeding a given distortion measure. Recently, with information bottleneck an alternative approach has been emerged in the field of machine learning. The fundamental idea is to include the original source into the problem setup when quantizing an observation variable and to use strictly information theoretic measures to design the quantizer. This paper yields an insight to this framework, discusses corresponding algorithms and their performance, and provides a new algorithmic approach of low complexity.**

I. Introduction

A fundamental task of any communication system is to quantize the noisy observation of the original source signal. The Rate-Distortion (RD) theory provides the minimal number of bits per symbol in order to represent the received signal without exceeding an upper-bound on a given distortion measure, e.g., the mean square error (MSE) between the quantizer input signal and its representative at the output [1]. Specifically, the Blahut-Arimoto algorithm determines the lowest achievable compression rate for a certain maximum tolerable distortion. The main drawbacks of this formulation are the lack of a systematic way to choose a proper distortion measure for any case of pertinence and the fact, that the stochastic relation between the noisy observation and the original data source is not considered. In [2], Tishby et al. have introduced the Information Bottleneck (IB) method for data compression. The central idea is to compress the observation such that the quantizer output preserves most of the information about the relevant variable, i.e., the original source. Furthermore, IB avoids the a priori specification of a distortion measure by considering the mutual information between the quantizer output and the original data source. In this fashion, the output of the quantizer becomes a compact representation of its input which is highly informative about the actual source of interest.

Clustering for dimensionality reduction is an important topic in learning theory, comprising a significant part of techniques dealing with the problem of unsupervised learning [3]. Along with the mentioned application which led to the introduction of the IB concept, similar quantization/compression problems arise in different aspects of data transmission like analog-to-digital converter (ADC) at receiver front-ends [4], implementation of discrete decoders for Low Density Parity Check (LDPC) codes [5] and many other potential cases.

In this paper we provide an overview of algorithmic implementations for IB-based quantization and compare their major performance metrics. To this end, the general IB framework is introduced in Section II and the principles of the most important, relevant algorithms are presented in Section III for arbitrary signal constellations. In Section IV the special case of binary input alphabet is considered. Section V is dedicated to the performance comparison of covered algorithms. Finally, the paper concludes by providing a summary of this study.

II. Information Bottleneck Method

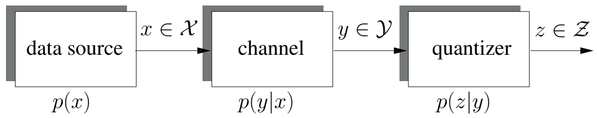

Fig. 1. General system model for the quantization of noisy observations

Fig. 1 shows the considered system model consisting of a data source, a transmission channel and a quantizer. Without loss of generality, we assume the random variable x with realizations $x \in \mathcal{X}$ following the probability mass function (pmf) $p(x)$ as a discrete memoryless source (DMS). The observation variable y with realizations $y \in \mathcal{Y}$ is the output of a discrete memoryless channel (DMC) characterized by its transition probability distribution $p(y|x)$. Furthermore, the random variable z with realizations $z \in \mathcal{Z}$ is the output of the quantizer block being characterized by the conditional distribution $p(z|y)$. Subsequently, $I(\mathsf{x};\mathsf{y}) = H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y})$ denotes the mutual information between x and y with the source entropy $H(\mathsf{x})$ and the conditional entropy $H(\mathsf{x}|\mathsf{y})$.

Given the joint probability distribution of the source and the channel output $p(x,y) = p(x)\,p(y|x)$ and assuming $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ to be a Markov chain, the quantizer should be designed such that the output z is a compact representation of the input y which is highly informative about x. Mathematically, the existent trade-off between the *compression rate*, $I(\mathsf{y};\mathsf{z})$, and the *relevant information*, $I(\mathsf{x};\mathsf{z})$, is established by the introduction of a non-negative Lagrange multiplier, β, in the design formulation. Hence, for an allowed number of quantizer output levels, n, the corresponding design problem follows as [2]:

$$p^\star(z|y) = \underset{p(z|y)}{\operatorname{argmin}}\ \frac{1}{\beta+1}\Big(I(\mathsf{y};\mathsf{z}) - \beta I(\mathsf{x};\mathsf{z})\Big) \quad \text{for } |\mathcal{Z}| \leq n\ . \quad (1)$$

ISBN 978-3-8007-4362-9

Note that the factor $\frac{1}{\beta+1}$ has been introduced here for the sake of mathematical clarity in the subsequent investigation of asymptotically small/large β without affecting the optimum mapping $p(z|y)$. To figure out the entity of this optimization task, two important questions must be answered at this point:

1) What is the event space of the mapping $p(z|y)$?
2) Is the objective function in (1) convex or concave over the corresponding event space?

To answer the first question, one notes that for each specific value y of the random variable y, the resultant $p(z|\mathsf{y}{=}y)$ is a $(|\mathcal{Z}|-1)$-dimensional probability simplex, since $\sum_{z\in\mathcal{Z}} p(\mathsf{z}{=}z|\mathsf{y}{=}y)=1$ holds. Hence, the overall event space of $p(z|y)$ is the product set of $|\mathcal{Y}|$ of such simplices, leading to a closed convex polytope in the $|\mathcal{Y}|\times(|\mathcal{Z}|-1)$ Euclidean space [6]. To answer the second question, we consider subsequently three different cases which cover the entire interval of allowed values of β, specifically two extreme cases of $\beta \to 0$ and $\beta \to \infty$ and the third case of non-zero finite values.

For $\beta \to 0$ the objective function in (1) reduces to the compression rate $I(\mathsf{y};\mathsf{z})$. As $I(\mathsf{y};\mathsf{z})$ is a convex function of $p(z|y)$ for fixed $p(y)$ [1], the optimization problem is *convex*. Any valid stochastic allocation of y to $|\mathcal{Z}|$ clusters that is repeated for all $y \in \mathcal{Y}$ is a solution, since in that fashion y and z become statistically independent and therefore the compression rate takes its global minimum value of $I(\mathsf{y};\mathsf{z}) = 0$. Obviously, $\beta \to 0$ is not a case of interest, as no relevant information is kept.

For $\beta \to \infty$, corresponding to the highest interest in keeping relevant information, the design problem (1) reduces to

$$p^\star(z|y) = \underset{p(z|y)}{\operatorname{argmax}}\, I(\mathsf{x};\mathsf{z}) \quad \text{for } |\mathcal{Z}| \le n \,. \tag{2}$$

For the present Markov chain $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ the conditional probability distributions $p(z|x)$ and $p(z|y)$ are connected by the affine relation $p(z|x) = \sum_{y\in\mathcal{Y}} p(z|y)p(y|x)$. Furthermore, it is known that any affine relationship preserves convexity. Therefore, as $I(\mathsf{x};\mathsf{z})$ is convex w.r.t. $p(z|x)$ for fixed $p(x)$, it is also a convex function of $p(z|y)$. Thus, the maximization in (2) is a *concave* optimization problem[1] [7]. Resorting to a well-known proposition in concave optimization theory [8] which asserts that a convex function $f : \mathcal{S} \to \mathbb{R}$ attains its global maximum over $\mathcal{S}$ at an extreme point of $\mathcal{S}$, one can deduce that there exists an optimal *deterministic* solution, i.e., $p(z|y) \in \{0, 1\}$, since extreme points of a convex polytope are its vertices. Each vertex of the event space of $p(z|y)$ corresponds to the product set of the vertices of its constituent probability simplices, leading to a deterministic mapping for each pair $(y, z) \in \mathcal{Y} \times \mathcal{Z}$. This is the principle behind most heuristics aiming to solve the problem (at least locally) in this case, as the naive exhaustive search over all $|\mathcal{Z}|^{|\mathcal{Y}|}$ vertices of the event space of the mapping $p(z|y)$ is obviously intractable for the relatively large cardinality of elements to be clustered.

[1]One must note that *concave optimization* is about finding the *maxima* of a *convex function* (∪) over a feasible region and, thus, is essentially different from *convex optimization* which searches for the *minima* of a *convex function*.

For non-zero finite values of β, the objective function in (1) is the sum of a convex ($\frac{1}{\beta+1}I(\mathsf{y};\mathsf{z})$) and a concave ($\frac{-\beta}{\beta+1}I(\mathsf{x};\mathsf{z})$) function of $p(z|y)$ which is neither in general. Hence, also the present optimization is neither of convex nor concave type.

It is noteworthy that the present constraint on the cardinality of representatives $|\mathcal{Z}| \le n$ in the problem statement always restricts the compression rate. So, even for $\beta \to \infty$, the compression rate is upper-bounded by $I(\mathsf{y};\mathsf{z}) \le \log_2(n)$ bits.

III. Algorithmic Approaches

In this section, we offer a review of IB-based algorithms for the scalar quantizer design in case of discrete random variable y. For continuous random variable y, interested readers are referred to [9]. In what follows, cluster, bin and class refer to the same concept and hence are used interchangeably.

A. Iterative Information Bottleneck (It-IB)

In order to solve (1), Tishby et al. derived the optimal quantizer by means of variational calculus [2]. Precisely, for a fixed value of β the quantizer $p(z|y)$ is a stationary point of the objective function in (1), if and only if

$$p(z|y) = \frac{p(z)}{\psi(y,\beta)} e^{-\beta D_{\mathrm{KL}}\left(p(x|y)\|p(x|z)\right)} \tag{3}$$

is fulfilled for all pairs $(y, z) \in \mathcal{Y} \times \mathcal{Z}$. The normalization function $\psi(y, \beta)$ ensures a valid distribution $p(z|y)$ for each $y \in \mathcal{Y}$ and $D_{\mathrm{KL}}(\cdot\,\|\,\cdot)$ is the Kullback-Leibler (KL) divergence[2]. The provided solution (3) has an implicit form, since the cluster probability $p(z)$ and the cluster representatives (in a conventional sense) $p(x|z)$ appearing on the right side of (3), depend on the quantizer $p(z|y)$ by

$$p(z) = \sum_{y\in\mathcal{Y}} p(y)p(z|y) \tag{4}$$

and

$$p(x|z) = \frac{1}{p(z)} \sum_{y\in\mathcal{Y}} p(x,y)p(z|y) \,. \tag{5}$$

The Iterative IB (It-IB) algorithm is initialized with a valid random mapping $p(z|y)$ and iterates over (4), (5) and (3), till a specific convergence criterion is met. It is noteworthy that for finite values of β, the resultant quantizer is of *stochastic* nature, i.e., each y is mapped to all clusters z with a certain probability $p(z|y)$. As the algorithm converges to a locally optimal solution, the iterative procedure is usually repeated for different initializations.

B. Agglomerative Information Bottleneck (Agg-IB)

The basic idea of the Agglomerative IB (Agg-IB) algorithm is to initialize z by the exact copy of y and then iteratively decrement the cardinality of representatives $|\mathcal{Z}|$ by merging two clusters in a greedy fashion till the allowed number of bins, n, is met [10]. Denoting the cluster created by merging

[2]The KL divergence is also known as relative entropy between two probability distributions $p(x)$ and $q(x)$ over the same event space $\mathcal{X}$ of the random variable x and is defined as $D_{\mathrm{KL}}\big(p(x)\|q(x)\big) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$ [1]. The mutual information of x and y is equal to the KL divergence $I(\mathsf{x};\mathsf{y}) = D_{\mathrm{KL}}\big(p(x,y)\|p(x)p(y)\big)$.

z_i and z_j by $\bar{z}$, the membership probability of the new class equals the sum of the corresponding probabilities [3]

$$p(\bar{z}) = p(z_i) + p(z_j) \tag{6}$$

with $p(x|\bar{z}) = \pi_i\, p(x|z_i) + \pi_j\, p(x|z_j)$ and $\Pi = \{\pi_i, \pi_j\} = \left\{\frac{p(z_i)}{p(\bar{z})}, \frac{p(z_j)}{p(\bar{z})}\right\}$. It is rather apparent that *minimizing* the objective function in (1) is equivalent to *maximizing* the functional $\mathcal{F} = I(\mathsf{x};\mathsf{z}) - \beta^{-1} I(\mathsf{y};\mathsf{z})$, since for constant β, it is only multiplied by a negative value $\frac{-(\beta+1)}{\beta}$. The merger cost, $\Delta\mathcal{F}$, defined as the difference between values of $\mathcal{F}$ before and after merging, is given by [3]

$$\Delta\mathcal{F}(z_i, z_j) = p(\bar{z}) \cdot \bar{d}(z_i, z_j) \tag{7}$$

where

$$\bar{d}(z_i,z_j) = D_{\text{JS}}^{\Pi}\big(p(x|z_i)\|p(x|z_j)\big) - \frac{1}{\beta} D_{\text{JS}}^{\Pi}\big(p(y|z_i)\|p(y|z_j)\big) \tag{8}$$

with the Jensen-Shannon (JS) divergence[3] $D_{\text{JS}}^{\Pi}(\cdot\,\|\,\cdot)$. It can be shown that the second divergence term in (8) simplifies to the binary entropy of Π in case of hard clustering. The greediness of the algorithm results from the fact that at each iteration, among all possible mergers, the one with the minimum cost is chosen. In this manner, it finds a quantizer mapping $p(z|y)$ which tries to directly maximize the functional $\mathcal{F}$.

C. Sequential Information Bottleneck (Seq-IB)

In [11] Slonim et al. presented a sequential algorithm for (1). Explicitly, this algorithm starts with a random deterministic classification, $p(z|y)$, with allowed number of bins, n, and then rearranges this mapping in an iterative fashion such that the functional $\mathcal{F}$ is maximized. Precisely, at each step, an element is drawn from its encompassing bin and considered as a singleton cluster. Then the merger cost of combining this cluster with all present classes is calculated using (7) and the one with minimum cost will be the new host for the dragged element. This procedure is repeated till a certain convergence criterion is met. Similar to the It-IB, to avoid getting stuck in bad local optima, the Sequential IB (Seq-IB) algorithm can be repeated for different initializations.

D. Deterministic Information Bottleneck (Det-IB)

Recently, the generalized objective function

$$\mathcal{L}_\alpha = H(\mathsf{z}) - \alpha\, H(\mathsf{z}|\mathsf{y}) - \beta\, I(\mathsf{x};\mathsf{z}) \tag{9}$$

with parameter $\alpha \in [0, 1]$ was introduced in [12] in order to find the optimal *deterministic* quantizer for the special case of $\alpha \to 0$. Note that, the stochastic nature of the solution provided by the It-IB algorithm stems from the presence of the term $H(\mathsf{z}|\mathsf{y})$ in the functional (1) in which $\alpha = 1$. With $\alpha \to 0$ the origin of stochasticity is suppressed leading to a deterministic solution $p(z|y)$ even for finite values of β. Again, for a fixed value of α the optimal quantizer is found by variational calculus

$$p(z|y) = \frac{1}{\psi(y, \alpha, \beta)} e^{\frac{1}{\alpha}\left(\log p(z) - \beta D_{\text{KL}}\left(p(x|y)\|p(x|z)\right)\right)} \tag{10}$$

where the normalization function $\psi(y, \alpha, \beta)$ ensures a valid distribution $p(z|y)$ for each $y \in \mathcal{Y}$. Obviously, decreasing the value of α translates into the asymptotically large growth of the power of the exponential function in (10). Hence, in the limit of $\alpha \to 0$, for each $y \in \mathcal{Y}$, the mapping $p(z|y)$ degenerates to a delta function, leading to a deterministic quantizer. Mathematically, the quantizer is given by $p(z|y) = \delta_{z,z^\star(y)}$ where δ denotes the Kronecker function and the optimum cluster $z^\star(y)$ is obtained as

$$z^\star(y) = \operatorname*{argmax}_{z}\Big(\log p(z) - \beta D_{\text{KL}}\big(p(x|y)\|p(x|z)\big)\Big) \,. \tag{11}$$

Like the It-IB (3), the derived solution (10) has an implicit form. Thus, to get the required quantizer $p(z|y)$, the Deterministic IB (Det-IB) algorithm is initialized by a valid random deterministic mapping $p(z|y)$ and iterates over equations (4), (5), and the provided mapping by (11), till a specific convergence criterion is met. Usually, the provided mapping does not use the entire allowed number of clusters, i.e., $|\mathcal{Z}| < n$, since the term $\log p(z)$ in (11) encourages the assignment of an element to already used bins.

E. KL-means Information Bottleneck (KL-means-IB)

As already discussed, for $\beta \to \infty$ the general IB problem (1) reduces to finding the quantizer $p(z|y)$ that maximizes the relevant information $I(\mathsf{x};\mathsf{z})$. Using the definition of mutual information and the Lemma for the difference of conditioned entropies provided in [13] we can write

$$\begin{aligned} I(\mathsf{x};\mathsf{z}) &= I(\mathsf{x};\mathsf{y}) - \big(H(\mathsf{x}|\mathsf{z}) - H(\mathsf{x}|\mathsf{y})\big) && \text{(12a)} \\ &= I(\mathsf{x};\mathsf{y}) - \mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{D_{\text{KL}}\big(p(x|\mathsf{y})\|p(x|\mathsf{z})\big)\right\} \,. && \text{(12b)} \end{aligned}$$

As $I(\mathsf{x};\mathsf{y})$ is fixed, the maximization of the relevant information $I(\mathsf{x};\mathsf{z})$ corresponds to the minimization of the average KL divergence $\mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{D_{\text{KL}}\big(p(x|\mathsf{y})\|p(x|\mathsf{z})\big)\right\}$. Similar to the Lloyd-Max algorithm [14], the KL-means-IB algorithm finds a locally optimal quantizer by alternating minimization in the mapping $p(z|y)$ (assignment step) and in the conditional probability distribution $p(x|z)$ (update step). The main difference here is, that the squared Euclidean norm used within the Lloyd-Max algorithm is substituted by the KL divergence being the proper distance measure for the IB setup. For initialization, the KL-means-IB algorithm selects randomly n points[4] $p(x|\mathsf{y} = y)$ corresponding to n different values of y as means of clusters. In the assignment step the points with smallest KL divergence to each mean are clustered in the same bin. Subsequently, in the update step the respective mean per cluster is recalculated as its center of mass [15]. This assignment and update procedure is repeated till a certain convergence criterion is met or a maximum number of iterations is reached.

[3]For two probability distributions $p(x)$ and $q(x)$ over the same event space $\mathcal{X}$ of the random variable x, the JS divergence is defined as $D_{\text{JS}}^{\Pi}\big(p(x)\|q(x)\big) = \pi_1 D_{\text{KL}}\big(p(x)\|r(x)\big) + \pi_2 D_{\text{KL}}\big(q(x)\|r(x)\big)$ where $\Pi = \{\pi_1, \pi_2\}$, $0 < \pi_1, \pi_2 < 1$, $\pi_1 + \pi_2 = 1$ and $r(x) = \pi_1 p(x) + \pi_2 q(x)$ [3].

[4]One may note that every conditional probability distribution $p(x|\mathsf{y} = y)$ can be regarded as a point in the space of dimension $|\mathcal{X}|$.

F. Channel-Optimized Information Bottleneck (Ch-Opt-IB)

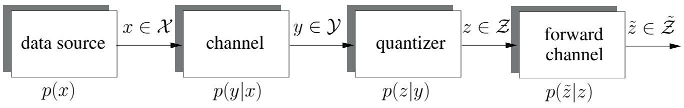

Fig. 2. The extended system model featuring forward channel

In [9], the quantizer design problem was extended by the transmission of the quantizer output signals z over a forward channel as depicted in Fig. 2. Denoting the output of the forward channel by $\tilde{\mathsf{z}}$ with realizations $\tilde{z} \in \tilde{\mathcal{Z}}$ and characterizing the transmission over the extra DMC by the transition probability distribution $p(\tilde{z}|z)$, the IB problem for $\beta \to \infty$ in (2) can be reformulated as

$$p^{\star}(z|y) = \underset{p(z|y)}{\operatorname{argmax}} \; I(\mathsf{x};\tilde{\mathsf{z}}) \;\; \text{for } |\mathcal{Z}| \le n \, . \tag{13}$$

This is again a *concave* optimization task for the Markov chain $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z} \leftrightarrow \tilde{\mathsf{z}}$, since $I(\mathsf{x};\tilde{\mathsf{z}})$ is a convex function of $p(\tilde{z}|x)$ for fixed $p(x)$ and, with a given forward channel $p(\tilde{z}|z)$, $p(\tilde{z}|x)$ and $p(z|y)$ are connected by an affine relation preserving convexity. Clearly, under the assumption of an error-free forward channel, the problem (13) equals (2). With

$$I(\mathsf{x};\tilde{\mathsf{z}}) = I(\mathsf{x};\mathsf{y}) - I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}}) \tag{14}$$

and fixed $I(\mathsf{x};\mathsf{y})$, the maximization of $I(\mathsf{x};\tilde{\mathsf{z}})$ in (13) equals the minimization of $I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}})$. Using the definition $C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) = D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big)$ the relation

$$I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}}) = \mathbb{E}_{\mathsf{y}}\big\{\mathbb{E}_{\tilde{\mathsf{z}}}\{C(\mathsf{y},\tilde{\mathsf{z}})|\mathsf{y}\}\big\} \tag{15}$$

has been derived in [9], where the conditional expectation calculates as

$$\mathbb{E}_{\tilde{\mathsf{z}}}\{C(\mathsf{y},\tilde{\mathsf{z}})|\mathsf{y}\} = \sum_{z\in\mathcal{Z}} p(z|y) \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)\, C(\mathsf{y}{=}y,\tilde{\mathsf{z}}{=}\tilde{z}) \, . \tag{16}$$

To minimize (16) for each $y \in \mathcal{Y}$, the quantizer mapping is chosen as $p(z|y) = \delta_{z,z^{\star}(y)}$ where the optimum cluster $z^{\star}(y)$ is obtained by

$$z^{\star}(y) = \underset{z}{\operatorname{argmin}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z) C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) \, . \tag{17}$$

Hence, the conditional distribution $p(\tilde{z}|y)$ of the combination of the quantizer and the forward channel calculates as

$$p(\tilde{z}|y) = \sum_{z\in\mathcal{Z}} p(\tilde{z}|z)p(z|y) = p(\tilde{z}|z^{\star}(y)) \, . \tag{18}$$

Apparently, in this fashion, the conditional mutual information (15) is minimized for a given $C(\mathsf{y},\tilde{\mathsf{z}})$. The basic idea of the Ch-Opt-IB algorithm is to adapt the iterative IB discussed in Section III-A to the current problem. Explicitly, it initializes to a random $C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z})$ for all $(y,\tilde{z}) \in \mathcal{Y}\times\tilde{\mathcal{Z}}$ and iterates over the modified versions of (4), (5) and (18) (substituting z by $\tilde{z}$), till a specific convergence criterion is met. Clearly, after each iteration $C(\mathsf{y},\tilde{\mathsf{z}})$ is updated accordingly. It is noteworthy, that the degenerated version of this algorithm assuming an ideal forward channel has already been proposed in [4].

IV. Algorithmic Approaches for Binary Alphabets

In this section, we focus on the special case of binary input alphabets, i.e., $|\mathcal{X}| = 2$, for which there exists an efficient algorithm that finds a globally optimum solution of the underlying *concave* optimization. Moreover, we discuss two other algorithmic approaches that are derived by modifying the Agg-IB and the Seq-IB from Section III leading to suboptimal solutions with reduced complexity.

A. Optimal Binary Algorithm (Opt-Binary)

In [7], Kurkoski and Yagi presented an algorithm which finds the optimal mapping $p(z|y)$ for $\beta \to \infty$. Based on the fundamental result on the existence of an optimal quantizer with all clusters being convex sets [16], their focal idea is to transform the quantization space suitably. For the binary input alphabet this transformation is achieved by mapping each received signal $y \in \mathcal{Y}$ to its corresponding conditional probability $p(x|y)$ for a specific value of x. Hence, after relabeling y values such that $p(\mathsf{x}=x|y)$ be in an ordered fashion, the algorithm finds an optimal mapping $p(z|y)$ for which clusters contain contiguous elements of y.

It is shown in [17] that the computational complexity load of the corresponding algorithm can be drastically reduced through careful definition of absolutely monotone matrices along with the application of the SMAWK algorithm [18] to find the maximum entry within each row. Moreover, [19] proposes an alternative problem formulation by finding the shortest path in a directed acyclic graph with a limited number of hops leading to a modified Bellman-Ford algorithm.

B. Modified Agglomerative IB (Mod-Agg-IB)

The central point in the Opt-Binary algorithm is that for relabeled y values with $p(\mathsf{x}=x|y)$ be in an ordered fashion, there exists an optimal quantizer for which clusters contain contiguous elements of y. We suggest to apply this proposition as a priori knowledge in order to decrease the complexity of the Agg-IB presented in Section III-B. Explicitly, in the novel Modified Agglomerative IB (Mod-Agg-IB) algorithm for each cluster the merger cost is calculated only w.r.t. adjacent bins, leading to a significant reduction in the computational complexity as demonstrated in Section V-A.

C. Modified Sequential IB (Mod-Seq-IB)

Similarly, a modified version of the Seq-IB discussed in Section III-C has been proposed in [5]. Specifically, the Mod-Seq-IB algorithm is initialized with a random mapping $p(z|y)$ with contiguous elements of y in each cluster. Subsequently, this natural ordering is kept by only inspecting the elements adjacent to the classification borders. The mentioned inspection is formalized through the introduction of two loops being responsible for right-to-left and left-to-right movements. A minor modification to the proposed algorithm which saves redundant calculations, is to enter the second loop only in case of no changes happening in the first one, since only in this instance, checking for left-to-right movement makes sense.

ISBN 978-3-8007-4362-9

V. Performance Evaluations

Subsequently, we investigate the performance and the complexity of discussed algorithms. First, the case of binary input alphabet is studied for which the Opt-Binary algorithm is compared with two suboptimal, low-complexity algorithms Mod-Agg-IB and Mod-Seq-IB. Afterwards, a non-binary input alphabet is assumed to assess the algorithms presented in Section III. We apply equiprobable BPSK ($x \in \{\pm 1\}$) and 4-ASK ($x \in \{\pm 1, \pm 3\}$) as input and assume AWGN channels with noise variance $\sigma_n^2 = 1$. Furthermore, to acquire the channel transition distribution $p(y|x)$, the continuous channel output is clipped at an amplitude of $3\sigma_n$ above the maximum input signal (i.e., 4 for BPSK and 6 for 4-ASK) and uniformly discretized to $|\mathcal{Y}| = 128$ values. In particular, we investigate the accuracy by the *mutual information loss* $\Delta I = I(\mathsf{x};\mathsf{y}) - I(\mathsf{x};\mathsf{z})$ and the complexity-precision trade-off by the corresponding *compression rate* $I(\mathsf{y};\mathsf{z})$ for different values of β over varying allowed number of clusters n. Finally, to get an impression about the complexity of the considered algorithms their average runtime per execution in MATLAB® is also provided.

A. Binary Input Alphabet

Fig. 3 a) visualizes the mutual information loss ΔI of the Mod-Agg-IB and the Mod-Seq-IB for varying β which controls the level of the appearing performance floor for these suboptimal algorithms. Please note, as the resultant mapping of the Mod-Seq-IB algorithm depends on the initialization, to achieve the corresponding curve, it has been run 10^5 times for each specific allowed number of bins n, with the best taken. It can be seen, that in case of $\beta \to \infty$ the loss in accuracy compared to the rather complex, Opt-Binary algorithm almost vanishes for both suboptimal approaches. Furthermore, the similar study with the conventional algorithms revealed, that the Agg-IB had the same performance as its modified counterpart, while the conventional Seq-IB performed worse compared to its modified version. The reason behind, can be attributed to the more suitable choice of initialization for the Mod-Seq-IB.

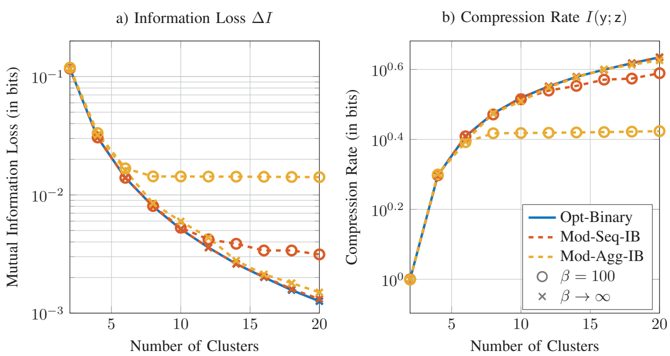

Fig. 3. a) Information loss ΔI and b) compression rate $I(\mathsf{y};\mathsf{z})$ for varying allowed number of bins n and binary input alphabet

For the same parameters the compression rates $I(\mathsf{y};\mathsf{z})$ are provided in Fig. 3 b). Considering both subfigures we can draw the conclusion, that the higher the accuracy, the higher is also the corresponding compression rate.

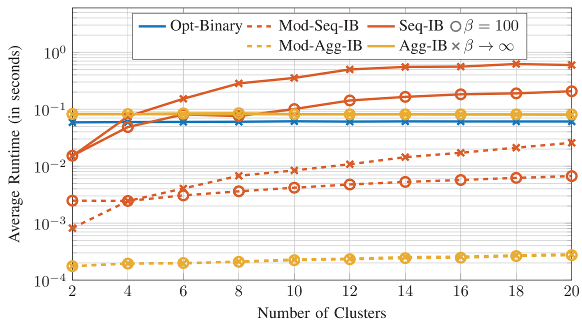

Fig. 4. Average runtime per execution for varying allowed number of bins n and binary input alphabet

Next, the average runtime per execution over varying allowed number of bins n is demonstrated in Fig. 4. To achieve so, for each algorithm, the corresponding arithmetic mean is calculated for 10^3 runs. It is readily observed, that both Mod-Agg-IB and Mod-Seq-IB end up to a significantly lower runtime compared to their conventional counterparts. Moreover, it can be deduced, that for $\beta \to \infty$ both suboptimal algorithms are a suitable substitute (having low-complexity and high-performance) for the Opt-Binary algorithm with high computational complexity. The time complexity of the Mod-Seq-IB algorithm increases with the allowed number of clusters n and is rather dependent on the parameter β. In contrast, the speed of convergence of the Mod-Agg-IB is nearly independent of β and the allowed number of classes.

Summarizing, in case of $\beta \to \infty$ the proposed Mod-Agg-IB algorithm shows a very good performance-complexity trade-off as it achieves high accuracy quantization with the least time complexity being independent of the number of bins n.

B. Non-Binary Input Alphabet

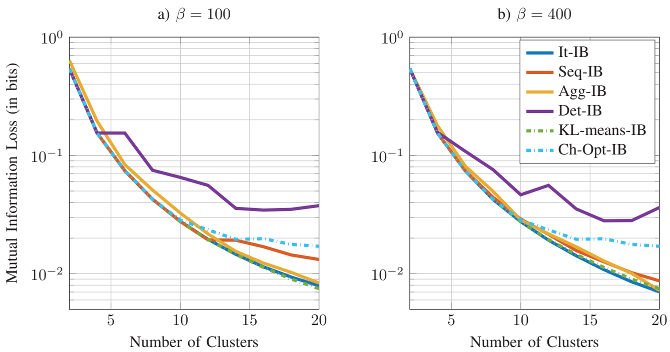

Fig. 5. Information loss ΔI for varying allowed number of bins n and 4-ASK input alphabet with a) $\beta = 100$ and b) $\beta = 400$

Fig. 5 shows the information loss ΔI of the algorithms presented in Section III. One may note, as the resultant mapping of all algorithms (except for the Agg-IB) depends on the initialization, to achieve the corresponding curves, they have been run 10^5 times, with the best taken. Except for the KL-means-IB and the Ch-Opt-IB (both only consider $\beta \to \infty$) one can observe, that the accuracy of all algorithms is improved by increasing β from 100 to 400. For a fair comparison with the KL-means-IB and the Ch-Opt-IB we concentrate subsequently on Fig. 5 b) with a relatively high value of β.

ISBN 978-3-8007-4362-9

First of all, the non-smooth behavior of the Det-IB is due to the fact that its provided mapping does not necessarily use the entire allowed number of clusters, i.e., $|Z| < n$. As an example, for $n = 12$ the used number of bins is smaller than the case of $n = 10$, leading to a coarser result. Furthermore, it can be seen that the It-IB and the KL-means-IB exhibit nearly the same performance over the entire range of allowed number of bins n. In addition, one notes that the Ch-Opt-IB also sweeps the corresponding curve of the It-IB for $n \leq 10$. The reason behind these observations is fully discussed in [20] where the asymptotic algorithmic equivalence of these algorithms is proven.

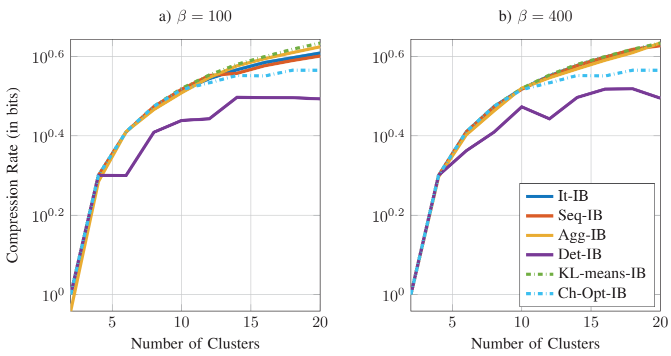

Fig. 6. Compression rate $I(\mathsf{y};\mathsf{z})$ for varying allowed number of bins n and 4-ASK input alphabet with a) $\beta = 100$ and b) $\beta = 400$

Fig. 6 displays the corresponding compression rates $I(\mathsf{y};\mathsf{z})$. Similar to the binary case we can observe, that in general, the lower the information loss introduced by quantization, the higher the corresponding compression rate.

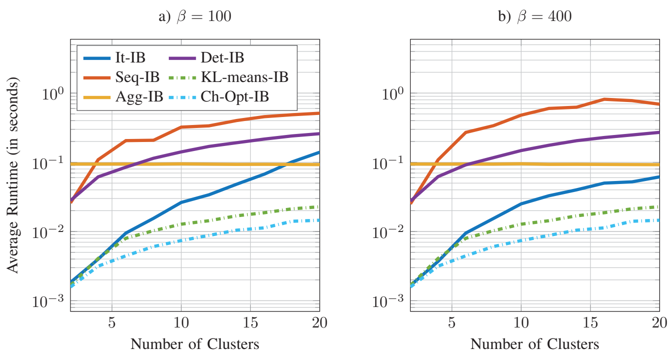

Fig. 7. Average runtime per execution for varying allowed number of bins n and 4-ASK input alphabet with a) $\beta = 100$ and b) $\beta = 400$

Finally, Fig. 7 visualizes the average runtime per execution of the algorithms indicating that the time complexity of the considered algorithms is nearly independent of β. Unlike the others, the complexity of the Agg-IB algorithm is further independent of the number of clusters. Moreover, it can be observed, that the Ch-Opt-IB algorithm exhibits the least time complexity. Furthermore, as suggested in Fig. 7 b), in general, the KL-means-IB exhibits lower time complexity compared to the It-IB. As a result, to avoid the present numerical instability, one may use the KL-means-IB algorithm as a good substitute of the It-IB algorithm for $\beta \to \infty$.

VI. Summary

In this study we discussed the general IB setup and provided an insight about the mathematical structure of the corresponding quantizer design problem. Then, we provided a succinct presentation of principles behind a group of algorithmic approaches to solve this task. For the covered algorithms we compared the accuracy, the compression rate, and the average runtime. It was demonstrated, that the proposed Modified Agglomerative IB algorithm provides a promising complexity-precision trade-off for binary input alphabets.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," in *37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Sep. 1999, p. 368–377.

[3] N. Slonim, "The Information Bottleneck: Theory and Applications," Ph.D. dissertation, Hebrew University of Jerusalem, Israel, 2002.

[4] G. C. Zeitler, "Low-Precision Quantizer Design for Communication Problems," Ph.D. dissertation, TU Munich, Germany, 2012.

[5] J. Lewandowsky and G. Bauch, "Trellis based Node Operations for LDPC Decoders from the Information Bottleneck Method," in *9th Int. Conference on Signal Processing and Communication Systems (ICSPCS 2015)*, Cairns, Australia, Dec. 2015.

[6] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "The Mathematical Structure of Information Bottleneck Methods," *Entropy*, vol. 14, no. 3, pp. 456–479, Mar. 2012.

[7] B. M. Kurkoski and H. Yagi, "Quantization of Binary-Input Discrete Memoryless Channels," *IEEE Trans. on Information Theory*, vol. 60, no. 8, pp. 4544–4552, Aug. 2014.

[8] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to Global Optimization*, 2nd ed. Springer Science & Business Media, 2000.

[9] A. Winkelbauer, "Blind Performance Estimation and Quantizer Design with Applications to Relay Networks," Ph.D. dissertation, TU Wien, Austria, 2014.

[10] N. Slonim and N. Tishby, "Agglomerative Information Bottleneck," in *Advances in Neural Information Processing Systems*, 1999, pp. 617–623.

[11] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised Document Classification Using Sequential Information Maximization," in *25th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, Aug. 2002, pp. 129–136.

[12] D. Strouse and D. Schwab, "The Deterministic Information Bottleneck," in *Conf. on Uncertainty in Artificial Intelligence*, New York, NY, USA, Jun. 2016.

[13] A. Zhang and B. M. Kurkoski, "Low-Complexity Quantization of Discrete Memoryless Channels," in *Int. Symposium on Information Theory and Its Applications (ISITA)*, Monterey, CA, USA, Oct. 2016.

[14] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[15] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, Oct. 2005.

[16] D. Burshtein, V. Della Pietra, D. Kanevsky, and A. Nadas, "Minimum Impurity Partitions," *The Annals of Statistics*, vol. 20, no. 3, pp. 1637–1646, Sep. 1992.

[17] K. Iwata and S. Ozawa, "Quantizer Design for Outputs of Binary-Input Discrete Memoryless Channels Using SMAWK Algorithm," in *IEEE Int. Symposium on Information Theory (ISIT)*, Honolulu, HI, USA, Jul. 2014.

[18] A. Aggarwal, M. M. Klawe, S. Moran, P. Shor, and R. Wilber, "Geometric Applications of a Matrix-Searching Algorithm," *Algorithmica*, vol. 2, no. 1-4, pp. 195–208, Nov. 1987.

[19] H. Vangala, E. Viterbo, and Y. Hong, "Quantization of Binary Input DMC at Optimal Mutual Information Using Constrained Shortest Path Problem," in *22nd Int. Conference on Telecommunications (ICT)*, Sydney, Australia, Apr. 2015, pp. 151–155.

[20] S. Hassanpor, D. Wübben, A. Dekorsy, and B. M. Kurkoski, "On the Relation Between the Asymptotic Performance of Different Algorithms for Information Bottleneck Framework," submitted to *IEEE Int. Conference on Communications (ICC),* Paris, France, May 2017.

